Article

Current Biology

The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes

Highlights

- We report the largest genomic dataset for African rice and its wild relative
- We infer the origin of African rice domestication in Northern Mali
- Rice domestication is associated with depletion of wild rice populations
- Convergent selection occurred during African and Asian rice domestications

Authors

Philippe Cubry, Christine Tranchant-Dubreuil, Anne-Céline Thuillet, ..., Olivier François, François Sabot, Yves Vigouroux

Correspondence

francois.sabot@ird.fr (F.S.), yves.vigouroux@ird.fr (Y.V.)

In Brief

The study of domestication provides insights about where and how major cultural transitions to agriculture have arisen. In this study, Cubry et al. document African rice (*Oryza glaberrima*) domestication, pinpointing its origin to the region of the Inner Niger Delta. A recent reduction of African rice cultivation is also evident.

Cubry et al., 2018, Current Biology *28*, 2274–2282
July 23, 2018
https://doi.org/10.1016/j.cub.2018.05.066

CellPress

CellPress

Current Biology
Article

The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes

Philippe Cubry,[1,12] Christine Tranchant-Dubreuil,[1,2,12] Anne-Céline Thuillet,[1,12] Cécile Monat,[1,2,12] Marie-Noelle Ndjiondjop,[3] Karine Labadie,[4,5,6] Corinne Cruaud,[4,5,6] Stefan Engelen,[4,5,6] Nora Scarcelli,[1] Bénédicte Rhoné,[1,7] Concetta Burgarella,[1] Christian Dupuy,[8] Pierre Larmande,[1,2,9] Patrick Wincker,[4,5,6] Olivier François,[10] François Sabot,[1,2,11,13,*] and Yves Vigouroux[1,11,14,15,*]

[1]Institut de Recherche pour le Développement, UMR DIADE, 911 Avenue Agropolis, 34394 Montpellier, France
[2]SouthGreen Development Platform, Agropolis Campus, Montpellier, France
[3]Africa Rice Center, AfricaRice, 01 B.P. 2031 Cotonou, Benin
[4]CEA, Institut de Biologie François Jacob, Genoscope, 2 Rue Gaston Crémieux, 91057 Evry, France
[5]CNRS, UMR 8030, CP5706, Evry, France
[6]Université d'Evry, UMR 8030, CP5706, Evry, France
[7]Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Lyon, France
[8]Institut des Mondes Africains (IMAF), Paris, France
[9]Institut de Biologie Computationnelle (IBC), Université Montpellier 2, 860 Rue St Priest, 34095 Montpellier Cedex 5, France
[10]Université Grenoble-Alpes, CNRS, UMR 5525 TIMC-IMAG, 38042 Grenoble, France
[11]Université de Montpellier, Place Eugène Bataillon, 34000 Montpellier, France
[12]These authors contributed equally
[13]Twitter: @francois_sabot
[14]Twitter: @YvesVigouroux
[15]Lead Contact
*Correspondence: francois.sabot@ird.fr (F.S.), yves.vigouroux@ird.fr (Y.V.)
https://doi.org/10.1016/j.cub.2018.05.066

SUMMARY

African rice (*Oryza glaberrima*) was domesticated independently from Asian rice. The geographical origin of its domestication remains elusive. Using 246 new whole-genome sequences, we inferred the cradle of its domestication to be in the Inner Niger Delta. Domestication was preceded by a sharp decline of most wild populations that started more than 10,000 years ago. The wild population collapse occurred during the drying of the Sahara. This finding supports the hypothesis that depletion of wild resources in the Sahara triggered African rice domestication. African rice cultivation strongly expanded 2,000 years ago. During the last 5 centuries, a sharp decline of its cultivation coincided with the introduction of Asian rice in Africa. A gene, *PROG1*, associated with an erect plant architecture phenotype, showed convergent selection in two rice cultivated species, *Oryza glaberrima* from Africa and *Oryza sativa* from Asia. In contrast, a shattering gene, *SH5*, showed selection signature during African rice domestication, but not during Asian rice domestication. Overall, our genomic data revealed a complex history of African rice domestication influenced by important climatic changes in the Saharan area, by the expansion of African agricultural society, and by recent replacement by another domesticated species.

INTRODUCTION

Agricultural societies emerged with the domestication of animals and plants. Studying the emergence of key crops can help us understand the historical processes that led to the development of such societies. In Africa, early agriculture was associated with the cultivation of African rice (*Oryza glaberrima* Steud.) and a few other cereals, including sorghum and pearl millet [1]. The domestication of African rice occurred independently from the domestication of Asian rice (*Oryza sativa* L.) [2, 3] and from a different wild progenitor, *Oryza barthii* A.Chev. (Figure 1). This wild African species diverged from the ancestor of Asian rice about one million years ago [4]. African rice domestication is widely acknowledged to have occurred in West Africa, but the circumstances leading to its domestication and its precise geographical origin are still debated [2, 3, 5]. We generated the most extensive genomic dataset to date and included samples from 163 domesticated *O. glaberrima* and 83 *O. barthii* individuals collected in the Sahel zone and East Africa (Table S1). Our new dataset pinpoints the cradle of rice domestication in the Inner Niger Delta. Our data support the hypothesis that climate hardship in the Sahara depleted wild resources and preceded domestication. Domestication of African rice was associated with an erect inflorescence phenotype as in Asian rice and selection for a loss of function on the exact same gene.

RESULTS

Genetic Diversity and Structure

The 163 domesticated and 83 wild relative individuals were fully sequenced with a high coverage (average 37x, range 20–55x).

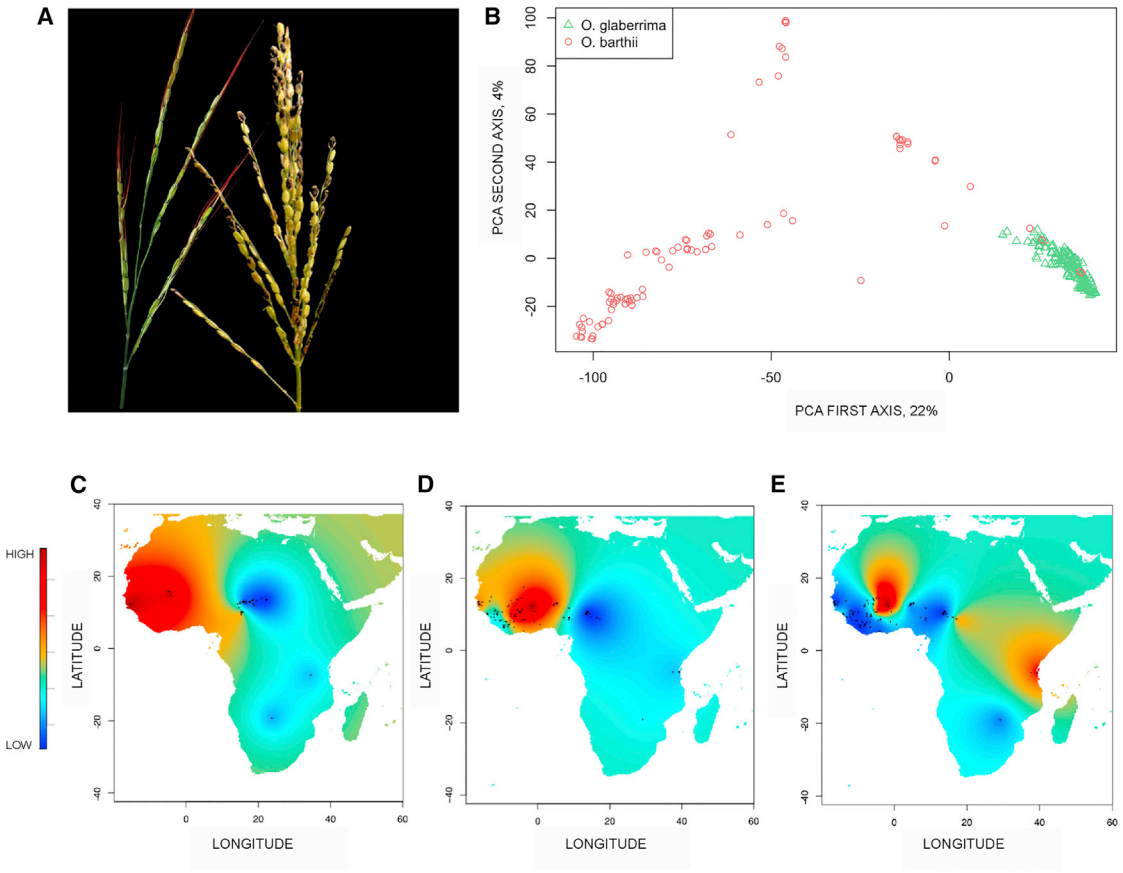

Figure 1. African Rice Diversity

(A) *O. barthii* (left) and *O. glaberrima* (right) panicles, 2017 © H. Adam.

(B) First and second axis of a principal-component analysis (PCA) including *O. glaberrima* samples (green triangles) and *O. barthii* samples (red circles).

(C) Geographical projection of the first PCA axis for the *O. barthii* samples. The high values on the map (red) indicated the proximity of the *O. barthii* samples with the *O. glaberrima* samples. Black dots represent coordinates of the wild samples.

(D and E) Geographical projection of the number of singletons (variants found in only one genotype) per sample for the chloroplastic (D) and nuclear (E) genomes for the *O. glaberrima* species. Black dots represent coordinates of the domesticated samples.

See also Figures S1 and S2 and Tables S1, S2, and S3.

A total of 3,051,681 nuclear SNPs (Figure S1 and Table S2) and 210 chloroplastic SNPs (Figure S1 and Table S3) were identified. Genetic diversity in domesticated African rice represented 54% of the genetic diversity of its wild relative (mean $\pi_{glaberrima} \approx 8.63 \times 10^{-4}$ and mean $\pi_{barthii} \approx 1.59 \times 10^{-3}$).

Principal-component analysis (PCA) based on nuclear SNP data showed strong genetic relatedness within the cultivated samples (Figure 1). The first PCA axis separated the wild samples from the domesticated samples. Clustering analysis also clearly separated wild and cultivated samples (Figure S2). The map of the first principal component provided evidence that the wild relatives closest to the domesticated samples originated from western Africa (Figure 1). We computed the geographic distribution of singletons defined as uniquely represented variants in the sample [6]. This statistic was evaluated for both species' nuclear and chloroplast genomes (Figure S2). This singleton statistic is correlated to π estimated on the same samples (Figure S2; $n = 14$, $r = 0.48$, $p < 0.05$). For *O. glaberrima*, the geographic distribution of chloroplastic singletons showed greater rare allele diversity in the central part of the Sahel area than elsewhere. The highlighted area encompassed present-day Mali, Ghana, Niger, Nigeria, Benin, and Togo (Figure 1). The geographic distribution of nuclear singletons highlighted a similar, even more restricted region in the central part of West Africa (Figure 1). An additional hotspot of rare alleles was observed in the eastern part of the study area for the nuclear singletons only. Nuclear rare allele diversity was lower in Guinea, Senegal, Ivory Coast, and Liberia, as well as east and south of Lake Chad (Figure 1). Excluding variants found in genes did not change the outcome of the analysis (Figure S2). The analysis for the wild species found a hotspot of rare alleles in East Africa for both chloroplastic and nuclear singletons supplemented by another hotspot of rare alleles in far western Africa for the nuclear set (Figure S2).

Past Population Effective Size History

Sequentially Markovian coalescent (SMC)-based methods pairwise SMC (PSMC) and multiple SMC (MSMC) [7, 8] were used to reconstruct the history of effective population size changes. Effective population size of *O. glaberrima* started to decline

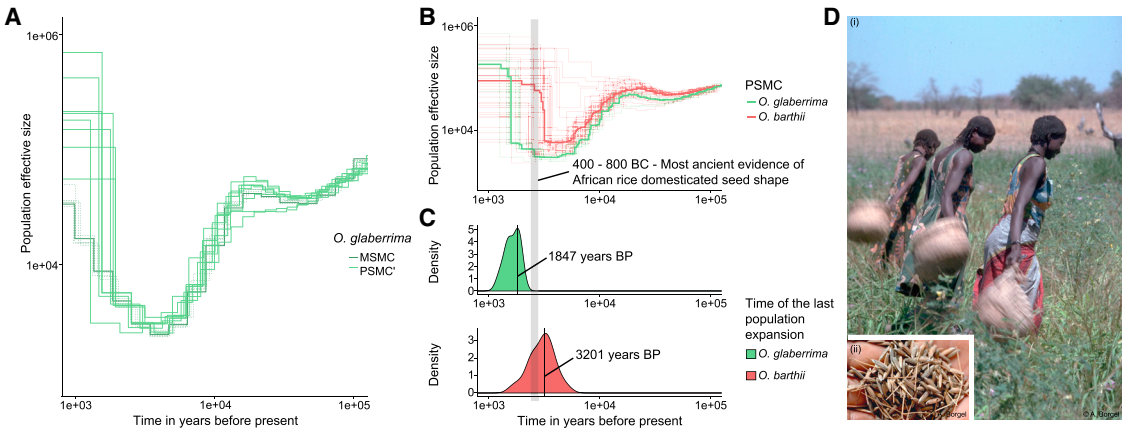

Figure 2. Past Effective Size Variation

(A) Past effective population size history of *O. glaberrima* assessed from the two coalescent approaches (PSMC, pale green, and MSMC, dark green). For MSMC, plain lines correspond to the result of the analysis, and dashed lines correspond to the corresponding bootstraps.

(B) PSMC analyses of both *O. glaberrima* and *O. barthii* show that the wild and cultivated populations shared a common history of reduced effective population size starting more than 10,000 years BP. Thick lines represent the median of the estimated sizes for *O. glaberrima* (green) and *O. barthii* (red), while thin lines are the PSMC results for each pair of considered genotypes (see STAR Methods for more details).

(C) Based on the last event of effective population size increase detected by PSMC in either *O. glaberrima* (green) or *O. barthii* (red), we estimated the timing of the end of the bottleneck, which showed a more ancient recovery for the wild than for the domesticated species.

(D) (Di) People harvesting wild rice fields using baskets in the region of Lake Chad, 1977 © A. Borgel. This activity was documented on the shores of Lake Chad and in the Ségou region of Mali (see text for references). (Dii) *O. barthii* seeds, Niger, Region of Diffa, 1976 © A. Borgel.

around 16,000–11,700 years before present (BP) and reached a minimum approximately 3,400 years BP (Figure 2). Effective population size experienced a strong increase approximately 1,850 years BP (mode = 1,847 years, SD = 438 years; Figure 2). The analysis of the wild populations (*O. barthii* accessions) also provided a signal of an ancient decrease that started more than 10,000 years ago for most of them. The effective population size reached a minimum about 3,800 years ago before it increased again after 3,200 years BP (mode = 3,201 years, SD = 1,655 years; Figure 2).

Next, we used SMC++ [9], a method tailored to handle large datasets and inference of demographic changes in very recent timescales. For older timescales, the same global trend as the one found with PSMC/MSMC analyses was recovered and included an important decrease of effective population size that began before 10,000 years BP (Figure 3). An effective population size of about 2,500 individuals was estimated for the bottleneck associated with the domestication. The cultivated population size increase began 3,500 years BP and was very slow at first. It remained low, with around 5,000 individuals, between 3,500 and 2,000 years BP. Domesticated African rice rose sharply after 2,000 years BP in perfect agreement with PSMC/MSMC results. African rice effective population size stabilized around 800 years ago, staying constant until 500 years BP. A sharp decrease was then observed, and population size dropped to half.

A Geographical Origin of Domestication in Northern Mali

The geographical origin of *O. glaberrima* in Africa was assessed using a spatially explicit coalescent simulation model [10]. Our spatial model simulated the diffusion of domesticated varieties throughout Africa (Table S4). We fitted our model using an approximate Bayesian computation (ABC) approach based on summary statistics derived from the site frequency spectrum (SFS) and from the empirical distribution of singletons in the sample. The analysis led to a probability map of the onset of culture expansion in northern Mali, with a maximum posterior probability at latitude 16.69°N and longitude 6.48°W (Figures 4 and S3 and Table S5). Predictive posterior checks indicated that the model simulated the observed data with good accuracy. All empirical summary statistics were found in the 95% credible intervals of their posterior predictive distributions (Figure S3).

Selection and Domestication

Selection during the domestication process is expected to decrease genetic diversity and to increase genetic differentiation compared to wild relatives. We used three complementary genome-wide methods to detect selection during the domestication process (Figure S4): (1) the detection of selective sweeps [11], (2) the genetic diversity ratio of the wild and cultivated species, and (3) the identification of loci with extreme genetic differentiation using F_{ST} calculations. The set of genes under selection (Data S1) were found to be significantly enriched in gene ontology (GO) terms related to growth and carbohydrate and protein binding (Table S6).

Among the most interesting candidates, the gene *prostrate growth 1* (*PROG1*; position 2,810,000–2,850,000 on chromosome 7, LOC_Os07 g05900) showed evidence of selection. This gene is also found under selection in *O. sativa* [12]. In Asian rice, *PROG1* is implicated in the transition between prostrate and erect growth [13, 14], a key architectural transition in the Asian rice domestication process. In Asian rice, a loss of function of this gene is associated with an erected phenotype and higher yield. The whole gene was fully deleted in the different assembled genomes of *O. glaberrima* [2, 15]. We confirmed its deletion in all the 162 cultivated samples. In contrast, *PROG1* was

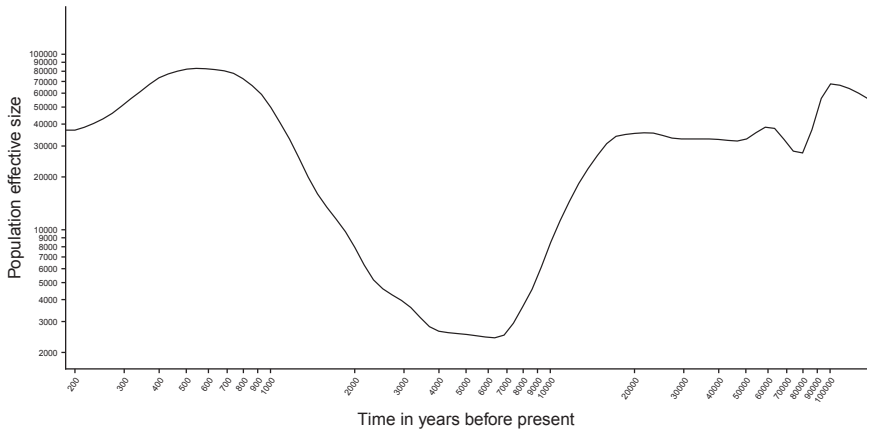

Figure 3. Inference of Population Effective Size History Using All Cultivated Genomes
Using all the cultivated individuals with the exception of those found in eastern Africa (Zimbabwe and Tanzania) and one exhibiting an odd number of singletons (namely MR) (Table S1), we inferred the effective population size history with the coalescent-based method SMC++.

deleted in only 6% of our wild samples, *O. barthii*. We validated (see STAR Methods) that the full *PROG1* deletion was associated with an erect phenotype (higher angle of branches) in (1) domesticated African rice and (2) wild individuals harboring the deletion ($p < 10^{-5}$; Figure 5). All wild individuals without the deletion presented a prostrate phenotype.

Dehiscence is also a key process of domestication [1, 16]. In Asian rice, two shattering genes were detected as strongly selected during the domestication process [16]: *SH4* and *OsSh1* (Figure 6). In our sample, none of these genes showed significant evidence of selection.

OsSh1 was previously described as missing in the assembly of *O. glaberrima* but described as present in the wild species [2]. In our dataset, *OsSh1* was absent in only 37% of the domesticated samples. Based on new *O. glaberrima* complete genome assemblies [15], we verified that the gene was present in some *O. glaberrima* genomes. If selection for the deleted allele occurred at this gene, it did not lead to its fixation. None of our wild *O. barthii* samples displayed deletion of *OsSh1*.

Another gene, *SH4/SH3/GL4*, was shown to be associated with the non-shattering phenotype and grain length in African rice [17, 18], and the potential causal mutation (a stop codon gain) occurred at a high frequency in *O. glaberrima*. In addition, the promoter of this gene, *pSH4*, was described as being under selection in African rice and expressed only in the wild species [2]. In our dataset, we found no evidence for strong selection at the *SH4* locus, nor at its promoter, despite the potential causal polymorphism of *SH4* found in approximately 86% of our *O. glaberrima* samples. The same polymorphism was present in only 2.5% of the *O. barthii* samples.

One gene of the shattering pathway, *SH5* (Figure 6), was found in a region under selection (position 22,310,001–22,360,000, chromosome 5, LOC_Os05 g38120). A deletion in the coding region of *SH5* was fixed in the cultivated species and was found in approximately half of our wild *O. barthii* samples.

DISCUSSION

A Geographic Origin Consistent with Both Ethnological and Archeological Studies

Wild accessions of West Africa were more closely related to the domesticated varieties than were the wild accessions of Chad or East Africa. Both for chloroplast and nuclear genomes, one major hotspot of rare variants was found in the Inner Niger Delta region in West Africa. We also detected an additional hotspot of rare variants in eastern Africa for nuclear diversity. This hotspot suggests that diversity in East African cultivated plants was enhanced by gene flow from some unknown relatives. As we did not find similar signatures at the chloroplastic level transmitted maternally, we hypothesized the occurrence of pollen gene flow with East African rice relatives (Figure S2). This rare variant hotspot is certainly also shaped by a recent population expansion [19] in eastern Africa (Figure S5).

We inferred the first probabilistic map of the geographic origin of domestication of African rice. The map pinpointed the north of the Inner Niger Delta as the most likely region, close to the most ancient archeological evidence of domesticated rice [20]. This center of origin had been previously postulated based on ethnobotanical and linguistic studies [21]. However, other authors hypothesized a more western origin for the domestication of *O. glaberrima* [2] or a diffuse origin across the Sahel [22]. Until now, genetic analysis of domestication origins primarily relied on comparisons between wild and cultivated species [23–25]. Our innovative approach allowed us to infer the origin of domestication statistically by analyzing the genetic diversity of the cultivated species. Such model-based inference allowed a direct test of hypotheses in domestication study (story testing) instead of trying to interpret the patterns of our data (storytelling), which can lead to implicit biases [26].

Expansion of Cultivated Rice Agriculture Accelerated 2,000 Years Ago

We found a long period of effective population size reduction beginning more than 10,000 years ago in both cultivated (*O. glaberrima*) and wild (*O. barthii*) species. This decrease was previously interpreted as a long protracted period of cultivation [3]. But the reduction of effective population size was also observed for the wild species at the continental scale, from Senegal to Chad (Figure 2). This strong decrease of wild population size partly reflects major environmental changes in the Sahara [27].

The decrease in effective population size reached a minimum around 3,400 years ago for both domesticated and wild species. The long decline in effective population size of the domesticated species reflects both environmental degradation affecting wild populations and a potential bottleneck associated with a protracted period of domestication. These two hypotheses are difficult to tease apart, as evidence for a wide use of wild plants in the Sahara was documented as early as 7,500 years cal BC [28]. Harvest of fully threshing wild rice was documented in a

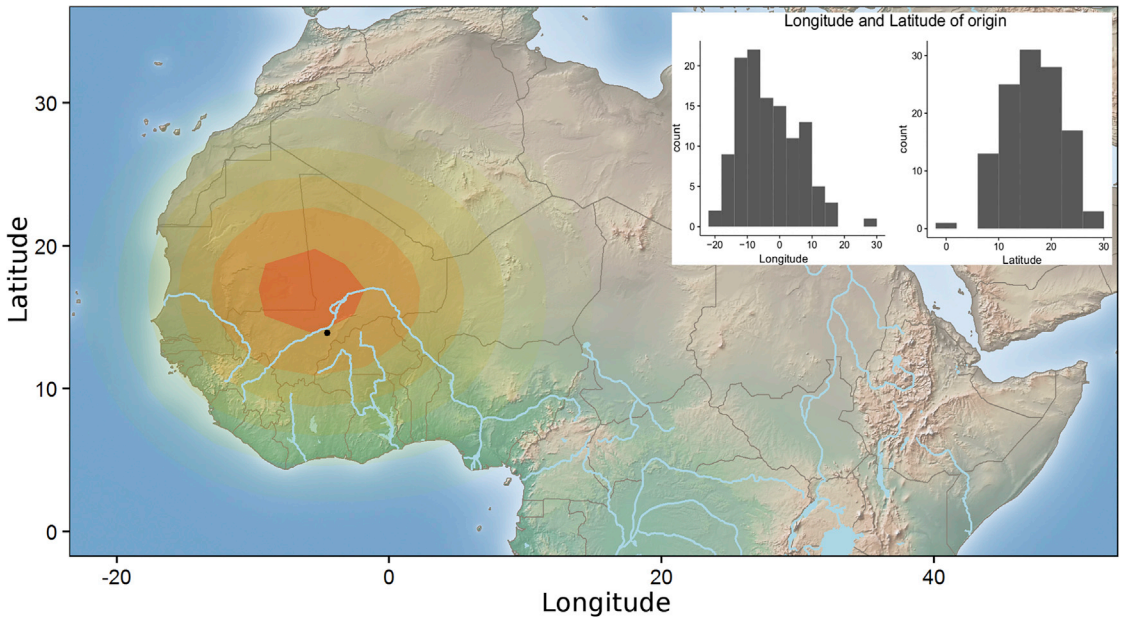

Figure 4. Inferred Origin of African Rice Expansion

Using spatially explicit coalescent simulations, we estimated the origin of the expansion of the domesticated samples. The dark dot indicates the location of the oldest known archeological record of grains with domesticated morphology.

See also Figures S3 and S5 and Tables S4 and S5.

temporary pond in northern Chad in the 1970s (Figure 2) and in Segou, Mali [29]. Wild rice was traditionally harvested in western Africa and might have been used as a resource in the past. A long protracted pre-domestication period of management and use of wild resources had also been suggested for the grapevine, *Vitis vinifera* ssp. *vinifera* [30]. Our results for African rice indicated that the wild resource might have become scarce due to less favorable climatic conditions, their overuse, or both of these processes. These results support the hypothesis that the drying of the Sahara directly or indirectly depleted wild resources and eventually led to domestication [31].

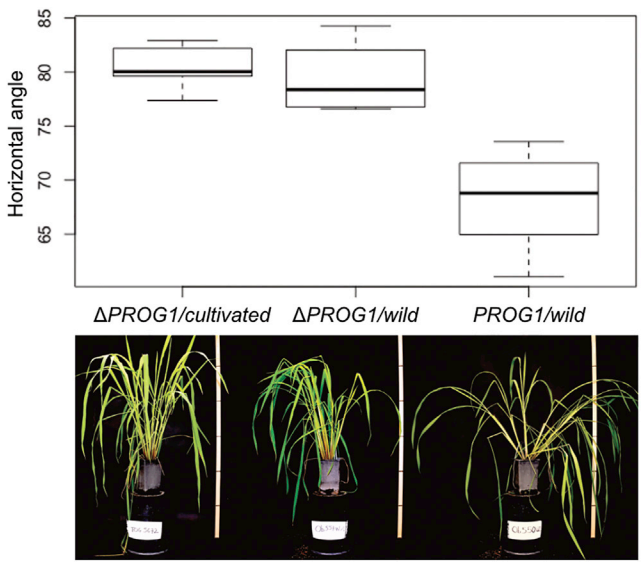

Figure 5. Erect Phenotype Is Associated with *PROG1* Deletion

The mean horizontal angle is higher for plants with the *PROG1* deletion ($p < 10^{-5}$), whatever the genetic background, domesticated or wild. Accessions photos (2017 © A.-C. Thuillet) illustrate each type: TOG5672, Ob557W1, and Ob550W1, respectively.

See also Figure S4.

For the wild species, a recovery of higher effective population size was observed around 3,200 years BP, an estimated date compatible with the end of the drying of the Sahara [27]. The cultivated history showed a longer period of low population effective size, with an expansion intensifying 2,000 years ago. During the time frame 3,200–2,000 years BP, archeological data suggest the presence of already fully domesticated grain [20].

A Recent Collapse of African Rice Agriculture

Inference based on all domesticated genomes [9] allowed us to highlight an unforeseen pattern of recent and strong decline of African rice agriculture. Effective population size was very high and was divided by two during the last 500 years. Although establishing a relationship between cultivated surfaces and effective population size is difficult, the result supports a drastic reduction of cultivation of African rice beginning 500 years ago. This decline is most probably explained by a switch from African rice to Asian rice cultivation by local farmers in West Africa. The oldest Asian rice remains found in Africa date back to 1,200–1,400 years ago on Comoros Island in the Indian Ocean [32]. Asian rice was massively introduced by the Portuguese in West Africa during the sixteenth century [33, 34]. Introduction of Asian rice intensified later, with a high number of Asian varieties introduced and adopted from the years 1870 to 1960 [21, 34]. In addition, African rice cultivation is more labor

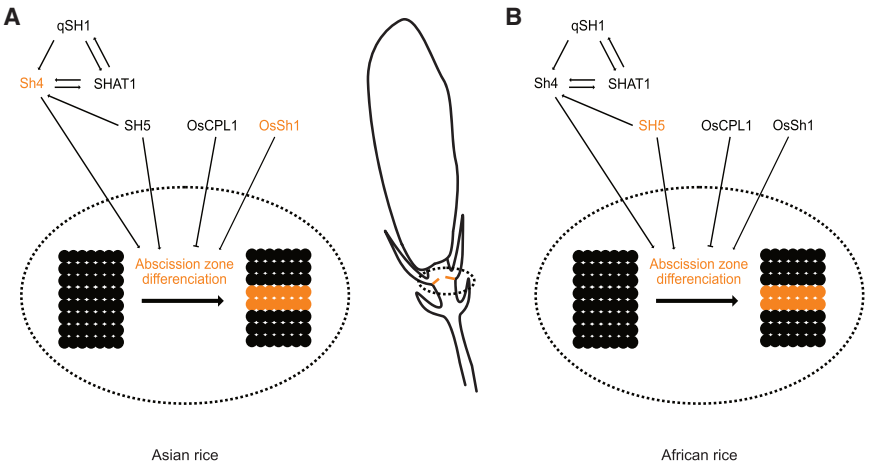

Figure 6. Regulatory Gene Network in Rice Involved in Abscission Zone Differentiation

(A and B) The currently accepted model of abscission describes a first step that corresponds to tissue differentiation defining the abscission zone (AZ). Genes showing evidence of selection in Asian rice (A) or in African rice (B) are colored in orange. Based on the review by Li and Olsen [16] and references inside, *O. sativa* (A) *Sh4* (LOC_Os04 g57530) carries a single non-synonymous substitution leading to partial function of the AZ, which is fixed in all cultivated Asian rice. *OsSh1*, a YABBY transcription factor (LOC_Os03 g44710), underlies a minor QTL in rice but was a target of selection during rice domestication. *qSH1* carries a causative mutation located 12 kb upstream of the BEL1-type homeobox gene (LOC_Os01 g62920), thus decreasing its expression level and interfering with the development of AZ in temperate japonica varieties only. For three other known shattering genes, *SHAT1* (LOC_Os04 g55560), *OsCPL1* (LOC_Os07 g10690), and *SH5* (LOC_Os05 g38120), no evidence of selection during the domestication process of Asian rice is known. In our study (B), evidence of selection during the domestication process of African rice was found only on *SH5*. In addition, *OsSh1* is absent from the genome of some individuals of our *O. glaberrima* sample, whereas it is present in all individuals of our sample of the wild relative.

See also Figure S4 and Table S6.

intensive [34], and the direct or indirect impact of slave trade on the West African agricultural economy might also have contributed to its decreased cultivation [35, 36].

Convergence in Selection during Asian and African Rice Domestication

We found a signature of selection on the gene *PROG1* during African rice domestication history. During domestication in Africa, the deletion of the gene was fully fixed in the domesticated compartment. The complete deletion of the gene is associated with a selection signature in African rice, whereas a loss of function mutation was also found to be under selection during Asian rice domestication [12–14]. In both Asian and African rices, samples harboring the *PROG1* deletion or loss of function show strong changes in their phenotype. We therefore found a signature of convergent selection between Asian and African rices at the gene *PROG1* during two independent domestication events from different wild progenitors.

Strong selection on shattering genes during domestication was different in Asian and African rice, but the gene *SH5* that we detected for African rice occurred in a similar abscission gene regulatory network (Figure 6). The associated deletion found in the coding region was annotated as a disruptive in-frame deletion and might reduce shattering. However, as the mutation is found at strong frequency in the wild sample, it alone might not explain the reduced shattering phenotype of African rice.

Another gene of the shattering pathway, *SH4*, which we did not detect as under selection, was previously found as implicated in the reduction of seed shattering in African rice [17, 18] and remains a strong candidate. In Asian domesticated rice, a functional mutation of *SH4* is implicated in reduced shattering and was found selected during domestication, a selection that led to its fixation in the cultivated species [37, 38]. Our study did not discover this gene because our detection of selection was tailored to detect strong positive selective sweeps [2]. Such an absence of selective signal was described in Asian rice with *qSH1*. This locus explained a large part of the non-shattering phenotype in some *O. sativa* ssp. *japonica* accessions [39]. Despite its demonstrated importance, selection at this locus was not detected in the *indica* varieties and was unclear for the *japonica* varieties [38]. Thus, other forms of selection, like soft sweep, might have shaped the dehiscence pathway in both Asian and African rice and might have not been detected by our approaches. Our results paved the way to a better understanding of the dehiscence phenotype in African rice, which will necessitate more quantitative trait loci (QTL) analyses in the near future.

Interestingly, we found major deletions associated with selection or phenotypic change like *PROG1* and *SH5*. Another example of such a process in African rice is the *OsSWEET14* gene [40]. There is a growing interest in pan and core genomes and their role in shaping diversity [41]. In Asian and African rices, it has been shown that structural differences may exist between individual genomes [15, 42]. With *PROG1*, we clearly found a gene exhibiting a polymorphism of presence/absence in the wild compartment, strongly selected in the domesticated samples for an interesting agricultural phenotype.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Samples used for genome re-sequencing experiment
  - Samples used for phenotypic experiment
- METHOD DETAILS
  - DNA extraction and sequencing
  - Variation calling and SNP annotation

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures, six tables, and one dataset and can be found with this article online at https://doi.org/10.1016/j.cub.2018.05.066.

ACKNOWLEDGMENTS

This work was supported by a grant from the France Génomique French National infrastructure and funded as part of "Investissement d'avenir" (ANR-10-INBS-09) and the IRIGIN project (http://irigin.org) to F.S. and an ANR grant (ANR-13-BSV7-0017) to Y.V. Y.V. is also supported by the Agropolis Resources Center for Crop Conservation, Adaptation and Diversity (ARCAD) with support from the European Union FEDER program and from the Agropolis Foundation (0900-001). P.C. and C.M. were supported by an ANR grant (AfriCrop project, ANR-13-BSV7-0017). C.M. and P.L. were also supported by a NUMEV labex grant (LandPanToggle, 2015-1-030-LARMANDE). Authors thank A. Ghesquiere, H. Chrestin, and S. Cheron-Perez for the maintenance of the African rices collection and J. Orjuela-Bouniol and I. Bourrié for their help in DNA extraction. We also want to thank Ndomassi Tando and the IRD itrop "Plantes Santé" bioinformatic platform for providing HPC resources and support for our research project.

AUTHORS CONTRIBUTIONS

Y.V. and F.S. managed the research. Y.V., F.S., and O.F. conceived the study and supervised the research. F.S., K.L., C.C., S.E., and P.W. managed and performed the data sequencing. C.T.-D., C.M., and F.S. performed and managed the bioinformatic analysis. P.C., Y.V., and O.F. performed and managed the model-based inference. A.-C.T., P.C., Y.V., O.F., C.T.-D., C.B., B.R., M.-N.N., N.S., C.D., and P.L. performed passport data collection, diversity analysis, structuration, and detection of selection. All authors contributed to drafting and writing the manuscript, which was led by Y.V. and F.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 12, 2018
Revised: April 10, 2018
Accepted: May 24, 2018
Published: July 5, 2018

REFERENCES

1. Doebley, J.F., Gaut, B.S., and Smith, B.D. (2006). The molecular genetics of crop domestication. Cell *127*, 1309–1321.
2. Wang, M., Yu, Y., Haberer, G., Marri, P.R., Fan, C., Goicoechea, J.L., Zuccolo, A., Song, X., Kudrna, D., Ammiraju, J.S.S., et al. (2014). The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication. Nat. Genet. *46*, 982–988.
3. Meyer, R.S., Choi, J.Y., Sanches, M., Plessis, A., Flowers, J.M., Amas, J., Dorph, K., Barretto, A., Gross, B., Fuller, D.Q., et al. (2016). Domestication history and geographical adaptation inferred from a SNP map of African rice. Nat. Genet. *48*, 1083–1088.
4. Huang, X., Zhao, Q., and Han, B. (2015). Comparative Population Genomics Reveals Strong Divergence and Infrequent Introgression between Asian and African Rice. Mol. Plant *8*, 958–960.
5. Bellwood, P. (2004). First Farmers: The Origins of Agricultural Societies (Wiley-Blackwell).
6. Cubry, P., Vigouroux, Y., and François, O. (2017). The Empirical Distribution of Singletons for Geographic Samples of DNA Sequences. Front. Genet. *8*, 139.
7. Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.
8. Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nat. Genet. *46*, 919–925.
9. Terhorst, J., Kamm, J.A., and Song, Y.S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat. Genet. *49*, 303–309.
10. Ray, N., Currat, M., Foll, M., and Excoffier, L. (2010). SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. Bioinformatics *26*, 2993–2994.
11. Pavlidis, P., Živkovic, D., Stamatakis, A., and Alachiotis, N. (2013). SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol. Biol. Evol. *30*, 2224–2234.
12. Huang, X., Kurata, N., Wei, X., Wang, Z.-X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H., Li, W., et al. (2012). A map of rice genome variation reveals the origin of cultivated rice. Nature *490*, 497–501.
13. Jin, J., Huang, W., Gao, J.-P., Yang, J., Shi, M., Zhu, M.-Z., Luo, D., and Lin, H.-X. (2008). Genetic control of rice plant architecture under domestication. Nat. Genet. *40*, 1365–1369.
14. Tan, L., Li, X., Liu, F., Sun, X., Li, C., Zhu, Z., Fu, Y., Cai, H., Wang, X., Xie, D., and Sun, C. (2008). Control of a key transition from prostrate to erect growth in rice domestication. Nat. Genet. *40*, 1360–1364.
15. Monat, C., Pera, B., Ndjiondjop, M.-N., Sow, M., Tranchant-Dubreuil, C., Bastianelli, L., Ghesquière, A., and Sabot, F. (2017). De Novo Assemblies of Three Oryza glaberrima Accessions Provide First Insights about Pan-Genome of African Rices. Genome Biol. Evol. *9*, 1–6.
16. Li, L.F., and Olsen, K.M. (2016). To Have and to Hold: Selection for Seed and Fruit Retention During Crop Domestication. Curr. Top. Dev. Biol *119*, 63–109.
17. Win, K.T., Yamagata, Y., Doi, K., Uyama, K., Nagai, Y., Toda, Y., Kani, T., Ashikari, M., Yasui, H., and Yoshimura, A. (2017). A single base change explains the independent origin of and selection for the nonshattering gene in African rice domestication. New Phytol. *213*, 1925–1935.
18. Wu, W., Liu, X., Wang, M., Meyer, R.S., Luo, X., Ndjiondjop, M.-N., Tan, L., Zhang, J., Wu, J., Cai, H., et al. (2017). A single-nucleotide polymorphism causes smaller grain size and loss of seed shattering during African rice domestication. Nat. Plants *3*.
19. Beissinger, T.M., Wang, L., Crosby, K., Durvasula, A., Hufford, M.B., and Ross-Ibarra, J. (2016). Recent demography drives changes in linked selection across the maize genome. Nat. Plants *2*.
20. Murray, S.S. (2004). Searching for the origins of African rice domestication. Antiquity *78*.
21. Portères, R. (1966). Les noms des Riz en Guinée (Fin). J. Agric. Trop. Bot. Appl. *13*, 641–700.
22. Richards, P. (1996). Agrarian creolization: the ethnobiology, history, culture and politics of West African rice. In Redefining Nature: Ecology, Culture and Domestication, R. F. Ellen and K. Fukui, eds. (Berg).
23. Matsuoka, Y., Vigouroux, Y., Goodman, M.M., Sanchez, G.J., Buckler, E., and Doebley, J. (2002). A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. USA *99*, 6080–6084.
24. Hufford, M.B., Martínez-Meyer, E., Gaut, B.S., Eguiarte, L.E., and Tenaillon, M.I. (2012). Inferences from the historical distribution of wild and domesticated maize provide ecological and evolutionary insight. PLoS ONE *7*, e47659.
25. Sedivy, E.J., Wu, F., and Hanzawa, Y. (2017). Soybean domestication: the origin, genetic architecture and molecular bases. New Phytol. *214*, 539–553.

26. Gerbault, P., Allaby, R.G., Boivin, N., Rudzinski, A., Grimaldi, I.M., Pires, J.C., Climer Vigueira, C., Dobney, K., Gremillion, K.J., Barton, L., et al. (2014). Storytelling and story testing in domestication. Proc. Natl. Acad. Sci. USA *111*, 6159–6164.
27. Kröpelin, S., Verschuren, D., Lézine, A.-M., Eggermont, H., Cocquyt, C., Francus, P., Cazet, J.-P., Fagot, M., Rumes, B., Russell, J.M., et al. (2008). Climate-driven ecosystem succession in the Sahara: the past 6000 years. Science *320*, 765–768.
28. Mercuri, A.M., Fornaciari, R., Gallinaro, M., Vanin, S., and di Lernia, S. (2018). Plant behaviour from human imprints and the cultivation of wild cereals in Holocene Sahara. Nat. Plants *4*, 71–81.
29. Morishima, H., Hinata, K., and Oka, H.-I. (1963). Comparison of Modes of Evolution of Cultivated Forms from Two Wild Rice Species, *Oryza breviligulata* and *O. perennis*. Evolution *17*, 170–181.
30. Zhou, Y., Massonnet, M., Sanjak, J.S., Cantu, D., and Gaut, B.S. (2017). Evolutionary genomics of grape (*Vitis vinifera* ssp. *vinifera*) domestication. Proc. Natl. Acad. Sci. USA *114*, 11715–11720.
31. Harlan, J.R., De Wet, J.M.J., and Stemler, A.B.L. (1976). Origins of African Plant Domestication (De Gruyter Mouton).
32. Crowther, A., Lucas, L., Helm, R., Horton, M., Shipton, C., Wright, H.T., Walshaw, S., Pawlowicz, M., Radimilahy, C., Douka, K., et al. (2016). Ancient crops provide first archaeological signature of the westward Austronesian expansion. Proc. Natl. Acad. Sci. USA *113*, 6635–6640.
33. Williamson, K. (1970). Some Food Plant Names in the Niger Delta. Int. J. Am. Linguist. *36*, 156–167.
34. Linares, O.F. (2002). African rice (Oryza glaberrima): history and future potential. Proc. Natl. Acad. Sci. USA *99*, 16360–16365.
35. Nunn, N. (2008). The Long-term Effects of Africa's Slave Trades. Q. J. Econ. *123*, 139–176.
36. Nunn, N., and Puga, D. (2010). Ruggedness: The Blessing of Bad Geography in Africa. Rev. Econ. Stat. *94*, 20–36.
37. Li, C., Zhou, A., and Sang, T. (2006). Rice domestication by reducing shattering. Science *311*, 1936–1939.
38. Zhang, L.-B., Zhu, Q., Wu, Z.-Q., Ross-Ibarra, J., Gaut, B.S., Ge, S., and Sang, T. (2009). Selection on grain shattering genes and rates of rice domestication. New Phytol. *184*, 708–720.
39. Konishi, S., Izawa, T., Lin, S.Y., Ebana, K., Fukuta, Y., Sasaki, T., and Yano, M. (2006). An SNP caused loss of seed shattering during rice domestication. Science *312*, 1392–1396.
40. Hutin, M., Sabot, F., Ghesquière, A., Koebnik, R., and Szurek, B. (2015). A knowledge-based molecular screen uncovers a broad-spectrum OsSWEET14 resistance allele to bacterial blight from wild rice. Plant J. *84*, 694–703.
41. Morgante, M., De Paoli, E., and Radovic, S. (2007). Transposable elements and the plant pan-genomes. Curr. Opin. Plant Biol. *10*, 149–155.
42. Schatz, M.C., Maron, L.G., Stein, J.C., Hernandez Wences, A., Gurtowski, J., Biggers, E., Lee, H., Kramer, M., Antoniou, E., Ghiban, E., et al. (2014). Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. Genome Biol. *15*, 506.
43. Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966–1967.
44. Monat, C., Tranchant-Dubreuil, C., Kougbeadjo, A., Farcy, C., Ortega-Abboud, E., Amanzougarene, S., Ravel, S., Agbessi, M., Orjuela-Bouniol, J., Summo, M., and Sabot, F. (2015). TOGGLE: toolbox for generic NGS analyses. BMC Bioinformatics *16*, 374.
45. Tranchant-Dubreuil, C., Ravel, S., Monat, C., Sarah, G., Diallo, A., Helou, L., Dereeper, A., Tando, N., Orjuela-Bouniol, J., and Sabot, F. (2018). TOGGLe, a flexible framework for easily building complex workflows and performing robust large-scale NGS analyses. bioRxiv. https://doi.org/10.1101/245480.
46. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12.
47. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
48. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
49. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
50. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.
51. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *11*, 11.10.1–11.10.33.
52. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.
53. Cingolani, P., Patel, V.M., Coon, M., Nguyen, T., Land, S.J., Ruden, D.M., and Lu, X. (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front. Genet. *3*, 35.
54. Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.
55. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576.
56. Leigh, J.W., and Bryant, D. (2015). POPART: full-feature software for haplotype network construction. Methods Ecol. Evol. *6*, 1110–1116.
57. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647–1649.
58. Sempéré, G., Philippe, F., Dereeper, A., Ruiz, M., Sarah, G., and Larmande, P. (2016). Gigwa-Genotype investigator for genome-wide analyses. Gigascience *5*, 25.
59. Milne, I., Shaw, P., Stephen, G., Bayer, M., Cardle, L., Thomas, W.T.B., Flavell, A.J., and Marshall, D. (2010). Flapjack–graphical genotype visualization. Bioinformatics *26*, 3133–3134.
60. Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J., and Holmes, I.H. (2009). JBrowse: a next-generation genome browser. Genome Res. *19*, 1630–1638.
61. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.
62. R Development Core Team (2016). R: A language and environment for statistical computing (R Foundation for Statistical Computing).
63. Whickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag).
64. Frichot, E., and François, O. (2015). LEA: An R package for landscape and ecological association studies. Methods Ecol. Evol. *6*, 925–929.

65. Csilléry, K., François, O., and Blum, M.G.B. (2012). abc: an R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. *3*, 475–479.

66. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

67. Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics *22*, 1600–1607.

68. Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., et al. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice (N. Y.) *6*, 4.

69. Orjuela, J., Sabot, F., Chéron, S., Vigouroux, Y., Adam, H., Chrestin, H., Sanni, K., Lorieux, M., and Ghesquière, A. (2014). An extensive analysis of the African rice genetic diversity through a global genotyping. Theor. Appl. Genet. *127*, 2211–2223.

70. Scarcelli, N., Mariac, C., Couvreur, T.L.P., Faye, A., Richard, D., Sabot, F., Berthouly-Salazar, C., and Vigouroux, Y. (2016). Intra-individual polymorphism in chloroplasts from NGS data: where does it come from and how to handle it? Mol. Ecol. Resour. *16*, 434–445.

71. Jukes, T.H., and Cantor, C.R. (1969). Evolution of protein molecules. In Mammalian protein metabolism, H.N. Munro, ed., pp. 133–182.

72. Hill, W.G., and Weir, B.S. (1988). Variances and covariances of squared linkage disequilibria in finite populations. Theor. Popul. Biol. *33*, 54–78.

73. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. Genetics *196*, 973–983.

74. Csilléry, K., Blum, M.G.B., Gaggiotti, O.E., and François, O. (2010). Approximate Bayesian Computation (ABC) in practice. Trends Ecol. Evol. *25*, 410–418.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
163 *O. glaberrima* leaf samples	IRD collection	See Table S1
83 *O. barthii* leaf samples	IRD collection	See Table S1
Critical Commercial Assays		
DNEasy Plant kit	QIAGEN	Cat#69181
NEBNext DNA Module	New England Biolabs	Cat#E7810S/L
AMPure XP	Beckman Coulter	Cat#A63882
Kapa Hifi Hotstart NGS library Amplification kit	Kapa Biosystems	Cat#KK2611
KAPA Library Quantification Kit	Kapa Biosystems	Cat#KK4824
Deposited Data		
fastq files	This study	SRA study: ERP023549 Accessions numbers ERX2068612-ERR2009097
SNPs	This study	http://gigwa.ird.fr:8080/gigwaV2/index.jsp
Experimental Models: Organisms/Strains		
6 *O. barthii* lines	IRD collection	See Table S1
3 *O. glaberrima* lines	IRD collection	See Table S1
Software and Algorithms		
Illumina Real-Time Analysis (RTA)	https://support.illumina.com/sequencing/sequencing_software/real-time_analysis_rta.html	RRID:SCR_014332
fastxtend	http://www.genoscope.cns.fr/externe/fastxtend/	N/A
SOAP aligner	http://soap.genomics.org.cn/soapaligner.html [43]	RRID:SCR_005503
TOGGLE v0.3	http://toggle.southgreen.fr/ [44, 45]	N/A
Cutadapt v1.8	https://cutadapt.readthedocs.io/en/stable/ [46]	RRID:SCR_011841
Burrow-Wheeler Aligner software v0.7.4	http://bio-bwa.sourceforge.net/ [47]	RRID:SCR_010910
SAMtools v0.1.18	http://samtools.sourceforge.net/ [48]	RRID:SCR_002105
GATK v3.3	https://software.broadinstitute.org/gatk/ [49–51]	RRID:SCR_001876
VCFtools v0.1.13	http://vcftools.sourceforge.net/ [52]	RRID:SCR_001235
SNPSIFT v4.2	http://snpeff.sourceforge.net/SnpSift.html [53]	RRID:SCR_015624
SNPeff v4.2	http://snpeff.sourceforge.net/ [54]	RRID:SCR_005191
VarScan v2.3.7	http://tvap.genome.wustl.edu/tools/varscan/ [55]	RRID:SCR_006849
popart	http://popart.otago.ac.nz/index.shtml [56]	N/A
Geneious Pro v4.7.6	https://www.geneious.com/ [57]	RRID:SCR_010519
Gigwa	http://www.southgreen.fr/content/gigwa [58]	N/A
FlapJack	https://ics.hutton.ac.uk/flapjack/ [59]	N/A
JBrowse	http://jbrowse.org/ [60]	RRID:SCR_001004
IGV	https://www.broadinstitute.org/igv/ [61]	RRID:SCR_011793
R v3.3.2	www.r-project.org [62]	RRID: SCR_001905
ggplot2 R package	http://ggplot2.tidyverse.org/ [63]	RRID:SCR_014601
LEA R package	https://github.com/bcm-uga/LEA [64]	N/A
MSMC2	https://github.com/stschiff/msmc2 [7, 8]	N/A
SMC++	https://github.com/popgenmethods/smcpp [9]	N/A
msmc_tools	https://github.com/stschiff/msmc-tools [8]	N/A
SPLATCHE2	http://www.splatche.com [10]	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
abc R package	https://cran.r-project.org/web/packages/abc/index.html [65]	N/A
SweeD v3.3.2	https://github.com/alachins/sweed [11]	N/A
BEDTools v2.2	https://github.com/arq5x/bedtools2 [66]	RRID:SCR_006646
TopGO R package	http://bioconductor.org/packages/release/bioc/html/topGO.html [67]	RRID:SCR_014798
Customized R scripts	https://github.com/Africrop/african_rice This study	N/A

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Yves Vigouroux (yves.vigouroux@ird.fr).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Samples used for genome re-sequencing experiment

We used 163 traditional cultivated varieties of the African rice, *O. glaberrima*, sampled during the period 1974 to 2005. The sampling covered West Africa exhaustively, supplemented by a few samples from East Africa. 83 wild relatives, *O. barthii*, were also sampled during the same period in Africa. We possess original record files or copies of the original description for the vast majority of the sampled individuals. From these original sampling sheets, we retrieved the latitude and longitude whenever possible (Table S1). We compared information from IRRI, AfricaRice, various databases and our own to ensure consistency in the coordinates. The discrepancies were resolved using the initial field records.

All samples were grown from seeds maintained in collection at IRD Montpellier, France, for three weeks before proceeding with DNA extraction. All varieties used for re-sequencing experiment were grown in green-house facilities of the IRD Montpellier, France, for three weeks before DNA extraction.

Samples used for phenotypic experiment

Three *O. glaberrima* (TOG6356, TOG5672 and TOG5307) and six *O. barthii* (Ob_602W, Ob_550W1, Ob_539W1, Ob_557W1, Ob_534A1 and OB_526W1) samples from the previous list were specifically used for a phenotypic experiment in IRD green-house facilities in Montpellier, France.

METHOD DETAILS

DNA extraction and sequencing

DNA sample preparation

Fresh leaves of 3-weeks old individuals were collected, frozen in liquid nitrogen, then milled using ceramic beads (QIAGEN, Germany) and stored at −80°C. DNA was then extracted from the frozen leaf powder using DNEasy Plant kit, as recommended by the supplier (QIAGEN, Germany). Free genomic DNA was collected in 100 μL of 0.1x TE, quantified using Nanodrop, and quality-checked using agarose gel (0.8%). 2 μg of high-quality DNA per sample was used for sequencing.

Sequencing

Libraries were prepared using the NEBNext DNA Module Products (New England Biolabs, MA, USA) with an 'on beads' protocol developed at the Genoscope. After gDNA fragmentation with the E210 Covaris instrument (Covaris, USA), end repair, A-tailing and ligation with adapted concentrations of Nextflex, DNA barcodes (Bioo Scientific, Austin, TX) were performed on the same AMPure XP beads that was used for the first purification after the end repair. After two consecutive 1x AMPure XP clean-ups, the ligated product was amplified by 12 cycles PCR using Kapa Hifi Hotstart NGS library Amplification kit (Kapa Biosystems, Wilmington, MA), followed by 0.6x AMPure XP purification. The library traces were validated on Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and quantified by qPCR using the KAPA Library Quantification Kit (Kapa Biosystems) on a MxPro instrument (Agilent Technologies, USA). The libraries were sequenced on Illumina HiSeq2000 or HiSeq4000 instruments (Illumina, USA), at 2x101 bp or 2x151 bp. respectively. Around 50 billion useful paired-end reads were obtained per run.

Quality control and initial treatments

Low quality clusters were filtered during the sequencing run using *Real Time Analysis* (*RTA*) software. Filtering steps were performed on all paired FASTQ files: Illumina adapters and primers were removed, nucleotides with quality value < 20 were trimmed from both ends, and the sequences between the second unknown nucleotide (N) and the end of the read were trimmed. Reads shorter than 30 nucleotides after trimming were discarded. These trimming steps were achieved using *fastxtend* (http://www.genoscope.cns.fr/externe/fastxtend/) – a software based on the *FASTX* library. The filtered reads and their mates mapped onto run quality control sequences (*PhiX* genome) were removed using *SOAP* aligner [43].

Variation calling and SNP annotation

Mapping and SNP calling

Read preprocessing and mapping was performed using the TOGGLE v0.3 pipeline [44, 45]. *Cutadapt* v1.8 [46] was used to trim the ends of the reads with low quality scores (–q 20) and only reads with a minimum length of 35 bp were kept. Paired-end reads were aligned to the Nipponbare reference genome (release 7.0 of the MSU Rice Genome Annotation Project/IRGSP 1.0 [68]) with the Burrow-Wheeler Aligner software v0.7.4 using the aln/sampe legacy approach [47]; standard options were used except for the edit distance set at 5 (-n option). The unmapped, abnormally mapped-in pair and non-unique reads were filtered with the SAMtools v0.1.18 software [48]. Local realignment was performed using GATK IndelRealigner [49–51].

SNP/indel were called using the GATK UnifiedGenotyper [49–51] v3.3. Raw SNPs were filtered with VariantFiltration from GATK according to the following criteria: quality score (QUAL) > 200, depth coverage (DP) > 10 and < 20,000, less than 3 SNPs into a window of 10 pb, Mapping Quality zero reads (MQ0)> = 4 and MQ0/(1.0xDP)>0.1.

Additional filters were applied with VCFtools v0.1.13 [52] and SNPSIFT v4.2 [53] to obtain high-quality SNPs: missing data < 10%, homozygous-variant called in more 90% of samples, only bi-allelic sites. SNP annotation was performed according to the Nipponbare reference genome using the SNPeff v4.2 tool [54]. The SNP location (exonic, intronic, intergenic, 5′UTRs or 3′UTRs, upstream and downstream) and functional annotation (synonymous, non-synonymous) were determined based on the genome annotation. These different filters enabled the retrieval of all 226 SNPs previously genotyped on the same set of accessions using Illumina VeraCode technology [69]. We then compared our SNP genotypes to that obtained from the VeraCode SNP chip in order to estimate the SNP genotyping error rate. We obtained 97% perfect matches between VeraCode genotypes and our dataset, the remaining 3% being mainly due to missing data in the VeraCode dataset. The error genotyping rate is consequently below 3%.

Validating SNP calling

To further validate the SNP calling, we used the opportunity of getting some duplicated individuals in our dataset. Three *O. barthii* accessions (*Ob_511W1*, *Ob_513W1* and *Ob_514G1*) were indeed sequenced twice (two sequencing runs from the same libraries). These duplicated individuals were submitted to the very same bioinformatic treatment. In order to get a confidence rate in SNP calling, we compared the alleles called at each positions (3 051 681) for the three pairs of duplicated individuals. We obtained a mean pairwise divergence rate between duplicates of 0.15%. Considering only singletons for these individuals (calculated excluding the duplicates), we found a mean pairwise divergence rate of 3.63%.

Chloroplast SNP calling and analyses

SNP calling was conducted according to Scarcelli et al. [70] on 246 samples (83 *O. barthii* and 163 *O. glaberrima*). We used SAMtools v1.1 mpileup [48] and VarScan v2.3.7 mpileup2snp [55] with a minimum variant allele frequency threshold (–min-var-freq) = 0.5 and minimum frequency to call homozygote (–min-freq-for-hom) = 0.5. The following descriptive statistics were calculated using VCFtools: mean depth, percentage of missing data and number of SNPs (Table S3). A Minimum Spanning Network was constructed using popart [56] (Figure S1). A Neighbor-Joining tree was constructed using Geneious Pro v4.7.6 [57] with 100 bootstraps and Jukes and Cantor [71] distance (Figure S1).

SNP database creation

We used the Gigwa [58] application to share our SNP datasets. This application provides an easy and intuitive way to explore large amounts of genotyping data by filtering it not only on the basis of variant features, including functional annotations, but also by genotype patterns. It is a web-based, platform-independent solution that feeds a MongoDB (https://www.mongodb.com/) NoSQL database with VCF [52] or HapMap files containing billions of genotypes, and provides a web interface to filter data in real time. Gigwa supplies the means to export filtered data in several popular formats, thus facilitating connectivity with many existing visualization engines such as FlapJack [59], JBrowse [60] and IGV [61].

QUANTIFICATION AND STATISTICAL ANALYSIS

Diversity and population genetics analysis

Linkage Disequilibrium

We used VCFtools to compute genome-wide pairwise linkage disequilibrium. We computed the haplotypic measure r^2 using a window of 1,000,000 base pairs for each SNP with a MAF of 10% and limiting the number of comparisons for each SNP to 1,000 (arguments–maf 0.1–hap-r2–ld-window-bp 1,000,000–ld-window 1,000). We then computed the mean value of r^2 for each distance in a range of 1 to 1,000,000 bp and plotted it against distance using a smoothing function (stat_smooth) from package ggplot2 [63] for R v3.3.2 [62]. We also derived the theoretical expectation [72] based on observed data (Figure S1).

Population structure

Principal Component Analysis (PCA, Figure 1) and a population structure analysis [73] (Figure S2) were performed using the R package LEA [64] (pca and snmf functions respectively, with K = 2 to 6 and 10 repetitions for snmf. Analyses were run on the whole sample including 83 O. barthii and 163 O. glaberrima and using a randomly chosen sample of 100,000 SNPs.

Data processing and site frequency spectrum computation

The data were processed using customized R scripts applied to the filtered VCF files generated from the bio-informatic pipeline described above. For subsequent analyses, each genotype was considered as fully inbred after randomly assigning one of the

two alleles of the few remaining heterozygote positions. This resulted in a single haplotype for each individual. A Site Frequency Spectrum (SFS) was computed from the allele counts for each position.

Geographical spanning of singletons

In a previous paper [6], we demonstrated that the use of singletons (alleles that were present in only one genotype) is a direct measure of genetic diversity using both theoretical derivations and simulations. We computed separate estimates of the number of singletons per genotype for the wild species *O. barthii* and for the cultivated species *O. glaberrima*. Singletons were calculated regardless of their status of derived or ancestral allele states. Estimates were obtained for each geo-referenced genotype and the results were interpolated on a geographical map. One cultivated individual (label MR, passport number 498) had twice more singletons than any other cultivated individual (Figure S2). For this individual, the singletons were not randomly distributed along the chromosomes as observed in the other cultivated individuals (Figure S2). Rather they clustered in specific genomic regions, which suggest introgression with distant wild relatives bringing new variants into the cultivated genome. Consequently, this individual was excluded from further analysis using singletons. To verify that our inference of genetic diversity based on the geographic distribution of singletons was not biased by potentially deleterious variants, we repeated the analysis on the cultivated species excluding all SNPs found in genes (based on the MSU7 genome annotation). The outcome of this analysis perfectly correlated with the analysis including SNPs found in genes (r^2 = 0.99, p value < 2.2e-16, Figure S2).

Regarding *O. barthii*, no genotypes exhibited a high excess or default of rare variants (Figure S2). We used a Kriging method to interpolate the distribution of singletons on a geographical map of Africa for both the chloroplast and nuclear genomes (Figure 1 for *O. glaberrima* and Figure S2 for *O. barthii* respectively).

Inference of effective population size history

We used the pairwise sequentially Markovian coalescent (PSMC) to infer historical changes in effective population size based on a single fully resequenced diploid individual [7] or two haploids. We also performed our analysis with a coalescent approach that allows up to 8 diploid individual genomes to be analyzed [8] (MSMC2, https://github.com/stschiff/msmc2). In addition, we implemented the SMC++ method [9] which can infer effective population size history from hundreds of individuals, and which is more powerful than PSMC/MSMC at recover history for very short time-scales.

PSMC/MSMC analysis

As we considered a selfing plant, we computed diploid genotypes by combining two haploid individuals as previously suggested [3, 19]. For *O. glaberrima*, we randomly selected 16 haplotypes to make diploid combinations. The resulting 8 diploids were analyzed either together (MSMC) or separately (PSMC). For the MSMC analysis, we additionally performed 6 bootstraps in order to evaluate confidence in the estimations. For *O. barthii*, we selected couples of haplotypes to be combined based on the ancestry coefficients obtained from the population genetic structure analysis with K = 6. We retained a minimum ancestry coefficient of 0.5 to assign a genotype to a given group. We then retained couples of genotypes to be combined within groups. Four combinations of chromosomes showed an extremely low number of SNP polymorphisms (less than 25% of the mean of other combinations), suggesting the two plants were related. We did not consider those four combinations during analyses. All together, we obtained 35 wild diploid individuals. We used the msmc_tools (https://github.com/stschiff/msmc-tools) approach to generate masks for positions that should not be mapped on the reference sequence. For PSMC/MSMC inferences, the scaling was made using a $6.5x10^{-9}$ mutation rate and a generation time of one year. We plotted the independent PSMC analyses for each considered pairs of cultivated and wild genotypes. To get an overview of the histories of both wild and cultivated populations, we calculated, with a 100 years interval, the median of the estimated effective sizes of wild or cultivated genotypes respectively (Figure 2). To assess the shared history between wild and cultivated species, we estimated the distribution of the date of the last population size increase in *O. barthii* and *O. glaberrima* (Figure 2). We estimated the mode of each distribution.

SMC++ analysis

We applied SMC++ approach to the cultivated species in order to get insight about the recent demographic events of the cultivated compartment. A single VCF containing pseudo-diploid genotypes generated from all available individuals with the exception of some outliers (genotypes from Tanzania and Zimbabwe as well as one from Senegal exhibiting an odd number of singletons (MR)) were generated for the cultivated species using the same approach as for PSMC. This resulted in a set of 79 pseudo-diploids for *O. glaberrima*. The masks used for PSMC/MSMC were also used for SMC++ in order to mask for very long runs of homozygosity that should lead to inaccurate estimations, especially in the recent past. We used 5 different distinguished lineages in order to get improved estimation as advised by the authors (https://github.com/popgenmethods/smcpp). We set the upper bound for the number of generations to estimate size history to 100,000 and let SMC++ calculate the lower bound based on a heuristic approach. We set the number of spline knots to be used in the internal representation of size history to 25. Estimation of size history was performed assuming a mutation rate of $6.5x10^{-9}$. All other parameters were set to their default values. Final representation of history of effective population size was made using a generation time of one year.

Inference of the geographic origin of the cultivated species

We used an Approximate Bayesian Computation (ABC) framework [74] to infer the geographical origin of the spread of the domesticated species *O. glaberrima* in Africa. ABC is a Monte Carlo approach that uses summary statistics calculated over the empirical data to estimate *a posteriori* distribution of model parameters that explain the observed data.

Simulation framework and model

We used SPLATCHE2 [10] to perform spatially-explicit simulations for the diffusion of the cultivated species in Africa. SPLATCHE2 enables users to model range expansions using non-equilibrium stepping-stone models. In those models, the geographical cells are colonized from their neighbors. We used information on the domestication process obtained from demographic inference with PSMC/MSMC and stairway plots to build a model. Following those results, we considered an initial bottleneck before the onset of range expansion. The time of occurrence, duration and intensity of this bottleneck were estimated by ABC. More specifically, we estimated the effective population size before expansion (Bott_Size, corresponding to population size during the bottleneck), the population size before the bottleneck (Anc_Size) and the duration of the bottleneck in generations (Bott_length). The priors on these parameters had uniform distributions. Our model included a recent change in effective population size that halved the carrying capacity of each local cell. The time of occurrence of this event (Rec_bott_time) was randomly drawn from a uniform distribution ranging in expansion time intervals. The expansion length (Exp_length) was randomly drawn from a uniform distribution, as well as the migration and growth rates.

We discretized the African continent as an array of 87 × 83 cells. The geographical origin of expansion was randomly chosen from a large geographical area encompassing the whole Sahel region. We set a carrying capacity (K) for each cell of the map, based on its eco-geographical region. For desert areas, the carrying capacity was fixed to K = 20, while for tropical rain forests, K was set to 10. For grasslands, savannas and tropical semi-deserts, K was randomly drawn from a uniform distribution between 30 and 150. A total of 100,000 independent loci were simulated. This analysis used a mutation rate high enough to allow mutations to occur during the simulation process ($u = 1x10^{-6}$). Individual genetic data were sampled at the same coordinates as those in the empirical dataset. The details for the priors used are provided in Table S4.

Summary statistics

We first defined geographical groups of genotypes using a kmeans clustering method. Using a rejection algorithm, we ran the kmeans algorithm until at least 12 groups including at least six genotypes were obtained. Four genotypes from the east region were grouped together (Figure S3). For those 13 groups, we computed the singleton statistics. These counted the average numbers of singleton per genotype in each of the 13 groups. We eventually normalized those quantities to obtain frequencies.

Next, we computed a SFS for the considered dataset. Because the SFS had high dimension with 111 georeferenced individuals, it was summarized using seven bins. The seven summary statistics grouped classes of the SFS as follows. We considered the following bins: group 1 (singletons), group 2 (bin of SNPs found two times), group 3 (bin of SNPs found three times), group 4 (bin of SNPs found 4 and 5 times), group 5 (bin of SNPs found from 6 to 10 times), group 6 (bin of SNPs found 11 to 20 times) and group 7 (bin of SNPs found more than 20 times). This partition resulted in seven summary statistics that described the SFS.

The 13 singleton statistics plus seven SFS bin statistics formed the vector of summary statistics under consideration in our ABC analysis. The same summary statistics were computed on the sample of georeferenced *O. glaberrima*. We replicated the analysis after excluding the four easternmost genotypes, and assessed the impact on the results.

ABC analysis and representation

We used the 'abc' R package [65] to perform ABC analysis. We used a neural network model with a tolerance rate of 0.1% and 500 neural networks to estimate posterior distributions for the parameters under consideration. The ABC analysis was made on both a 12 and a 13-groups sample, including or excluding the four easternmost samples in the analysis (Figure S3, Table S5). We represented a posterior distribution of the coordinates of the onset of expansion using a two-dimensional kernel density estimation. We used 100 grid points in each direction and a bandwidth of 30 to perform this analysis (Figure 4 and Figure S3 for the 12 and 13 groups models respectively).

Posterior Predictive Checks

To assess performance of the model in simulating datasets close to the observed ones, we used the posterior distribution of the parameters to feed a new set of simulations. We performed 500 additional simulations and calculated 95% confidence intervals of the summary statistics after Bonferroni's correction for multiple tests on the simulated data. For the 12 groups model, all observed summary statistics fell into the confidence intervals of the posterior predictive distribution, indicating that our model was able to predict our summary statistics accurately. Including the four easternmost genotypes, i.e., using 13 groups in the model, we correctly predicted 19/20 summary statistics. We represented these outcomes graphically (Figure S3). Posterior predictive checks indicated a better fit of coalescent models when the four eastern genotypes were not included. Nonetheless, the inclusion of those genomes in the ABC analysis pinpointed the same geographical origin as that obtained by excluding them (Figure 4 and Figure S3).

Detection of selection

Detection of selection during the domestication process was investigated using three methods (Figure S5). We used a likelihood ratio test for the detection of selective sweeps, based on SFS calculations and run with the software SweeD v3.3.2 [11]. SweeD was run on the *O. barthii* samples and on the *O. glaberrima* samples with a grid parameter of 10,000. Only peaks that were specific to the *O. glaberrima* samples were considered for further analysis. Composite-Likelihood Ratio (CLR) peak positions with distance less than 100kb from each other were considered as potentially resulting from similar selection events, and were grouped into common windows, including 50kb before and after the most extreme peak positions of the window. The two other methods were the π_i Ratio, calculated as π_i *O. barthii* / π_i *O. glaberrima* and the F_{ST} calculation of Weir and Cockerham between both species. The π_i and F_{ST} values were calculated using VCFtools, with a sliding window of 50,000bp every 10,000bp. For each of those three methods, the 1% most extreme values were considered to identify for each retained peak the window corresponding to the highest value

as a candidate genomic region for selection (Data S1). The treatment of output data from SweeD and VCFtools to obtain a list of candidate regions was performed using a customized script running under the R software (provided at https://github.com/Africrop/african_rice).

The list of candidate genomic regions for selection was crossed with the annotated genome of rice (Release 7.0 of the Rice Genome Annotation Project) using BEDTools v2.2 [66] (windowBed command, default parameters) to identify locus annotation.

Gene Ontology (GO) terms analysis

GO terms were associated with our lists of loci using a customized R script. Enrichment tests for GO biological process, cellular component and molecular function terms were performed using the Fisher exact test implemented in the R package TopGO [67].

Validating PROG1 *deletion in* O. glaberrima

To validate the absence of *PROG1*, we looked for its genomic sequence in three different *O. glaberrima* assemblies from [15] through BLAST, as well as on a set of PacificBiosciences long reads from the CG14 and TOG 5681 *O. glaberrima* varieties [40] (12x depth). We could not find the corresponding genomic sequence in any of these datasets, and thus concluded that *PROG1* were absent in the African rice genome. Using the mapping data of the whole set of *O. glaberrima*, as well as the different assemblies available for *O. glaberrima* [2, 15], we were able to fringe the deletion of the *PROG1* locus on Chromosome 7. The whole mapping of the 182 individuals shown a clear 5′ border at position 2,831,302, the 3′ position being little less precise between 2,843,710 and 2,843,714. On assemblies, the 5′ border is exactly the same, and the 3′ one is in the middle position, i.e., 2,843,712. The deletion is therefore similar across the entire studied individuals, wild and cultivated. This result suggests the same genetic variant might be shared by all the cultivated samples.

Phenotyping experiment

To test for the effect of *PROG1* presence or absence on plant growth (prostrate versus erect growth respectively), we designed a simple phenotyping experiment. This experiment was conducted in a green-house in Montpellier, France. We sowed three wild accessions possessing the *PROG1* gene (Ob_602W, Ob_550W1 and Ob_539W1), and three wild accessions (Ob_557W1, Ob_534A1 and OB_526W1) and three cultivated accessions (TOG6356, TOG5672 and TOG5307) deleted for this gene. Two repetitions of each accession were phenotyped. One accession (Ob_534A1) did not properly germinate and consequently was not phenotyped. Phenotyping measurements were performed 62 days after sowing (DAS). We measured the horizontal angle of each tiller and calculated a mean angle for each plant. ANOVAs were performed to test for the effect of wild or cultivated genetic background and presence/absence of *PROG1* gene in the plant architecture.

DATA AND SOFTWARE AVAILABILITY

All fastq files have been deposited in the Sequence Read Archive (SRA) under ID codes ERP023549 for the study, PRJEB21312 for the project and accessions numbers ERX2068612–ERR2009097. Nuclear SNPs (*vcf.* format) were deposited in a Gigwa database and are accessible at: http://gigwa.ird.fr:8080/gigwaV2/index.jsp.

Customized R scripts for the different analysis led in this study are provided in a github repository accessible at: https://github.com/Africrop/african_rice.